Thesauri Usage in Information Retrieval Systems: Example of LISTA and ERIC Database Thesaurus

Kristina Feldvari
Departmant of Information Sciences, Faculty of Philosphy in Osijek
Lorenza Jägera 9, Osijek, Croatia
kfeldvari@ffos.hr

Summary

This paper offers some thoughts on the usage of thesaurus in information retrieval with special reference to information retrieval systems like databases. A thesaurus is an example of controlled vocabulary and an important aid in subject analysis. Controlled vocabularies are used to describe the subject within knowledge organization systems, where the sole purpose of a vocabulary control is to achieve a consistency of subject description and facilitation of information retrieval. We survey some approaches to this question in literature and give two examples of usage of thesaurus in the following databases: the Thesaurus of ERIC Descriptors and LISTA thesaurus. These thesauri are described along with their functions and display in database.

Key words: thesaurus, information retrieval systems, Thesaurus of ERIC Descriptors, LISTA thesaurus

Introduction

Main research purpose is presentation of thesauri, their function and role in facilitation of information retrieval as well as their function and importance in information retrieval systems. A thesaurus is an example of controlled vocabulary and an important aid in subject analysis. It controls the vocabulary and is formed in a way that facilitates seeking and marking within a specific subject area. It actually has a place at both ends of the information access process, at both storage and retrieval.
Main research questions are: 1) How do IR systems (data bases) use thesauri?
2) Which functions do thesauri support? 3) How are thesauri displayed in data bases? We used the following methods: literature overview, analysis and the comparison of the data.

Purposes of controlled vocabularies

A keyword search for information on a particular subject performed on the World Wide Web may retrieve thousands of irrelevant documents[1]. According

[1] Svenonius, Elaine. Intelektualne osnove organizacije informacija. Lokve: Benja, 2005. Page 125

to Lancaster the major defect of the Internet as an information source, apart from its size, is the fact that it lacks any form of quality control.[2] We can try to solve this problem by using a subject language that incorporates measures designed to improve retrieval of the desired information. Usage of subject languages to retrieve information provides a value-added quality, which, in the case of highly refined languages, can transform information into knowledge.[3]
A subject language is used to describe what the document is about. The main purpose it serves are primary those of the collocation of documents that have the same information content and the navigation of the users. To achieve the collocation objective, the language must be designed so as to facilitate the retrieval of all and only relevant documents[4]. This is estimated by the twin measures of precision[5] and recall[6].
We can name five main purposes that controlled vocabularies serve:
1) Translation: provide a means for converting the natural language of authors, indexers, and users into a vocabulary that can be used for indexing and retrieval. 2) Consistency: promote uniformity in term format and in the assignment of terms. 3) Indication of relationships: Indicate semantic relationships among terms. 4) Label and browse: provide consistent and clear hierarchies in a navigation system to help users locate desired content objects. 5) Retrieval: serve as a searching aid in locating content objects.[7]
The last purpose is the our research question and will be elaborated in the rest of this paper.

What is thesaurus?

Definition

Librarian Lexicon defines thesaurus as a vocabulary of key words, i.e., a standardized set of terms and phrases authorized for use in an indexing system to describe a subject area or information domain.[8] If we take a look at definitions of some authors, e.g. D. Bawden we can notice that he offered definition of the-

[2] Lancaster, F. W. Do indexing and abstracting have a future? // Anales de documentation. 2003, 6; page 137

[3] Svenonius, Elaine. Ibid. Page 125

[4] Ibid, page 126

[5] Recall (R) is the proportion of relevant material retrieved and precision (P) is the proportion of retrieved material that is relevant.

[6] Lancaster, F.W. Indexing and abstracting in theory and practice. Compaing Illions: University of Illinois, 1998. Page 3-4

[7] ANSI/NISO Z39.19-2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. 2005. http://eric.de.gov/ERICWebPortal/resources/html/help/Z39-19-2005.pdf (2009-07-22)

[8] Bibliotekarski leksikon. Beograd: Nolit, 1984. Page 186

saurus purpose. He accents that in information retrieval thesaurus limits and controls the diversity of natural languages by offering an expression that should be use for each concept[9] while M.L. Nielsen says that thesaurus is a well-known and well-established tool in information retrieval that is used to guide indexing and retrieval based on controlled as well as natural language indexing.[10]

Thesaurus deficiencies- user comprehension and usage

J.Greenberg in her article stresses out three questions when she talks about user thesaurus comprehension: thesaurus interface design, processing options and end-user warrant. Current thesaurus-supported systems often fail to adequately highlight the thesaurus search option. Information systems may include the word ¨thesaurus¨ on a navigation bar or as a hypertext button, but the explanation of how this feature can assist with the selection of search terms may be hidden. Additionally, systems that include the thesaurus often provide confusing interfaces. They use thesaural identifiers like ¨BT¨ and ¨NT¨ which may not be clear to a user. In this study it is also concluded that if we give a basic thesaurus introduction, users will indicate a desire to employ these tools and also that users favor either interactive or a combination of automatic and interactive thesaurus processing compared to completely automatic processing.[11]

Second question ¨end-user warrant¨ is explained by M. Bates. She proposes that matching and lead-in terminology should be made available for information searchers to help them in their search process. Such an end-user thesaurus would recognize the many variants, informal terms and other terms that users actually input when searching. The thesaurus would be designed to link directly with whatever database the searcher wanted to use, so that the searcher could be led to the ¨legitimate¨ indexing terms.[12]

Thesaurus and IR systems

Information retrieval is defined as the process of searching a collection of documents, using the term document in its widest sense, in order to identify those documents which deal with a specific subject. The success is determined

[9] Bawden, David. Tezaurusi: nova postignuća. // Vjesnik bibliotekara Hrvatske. 44 (2001), 1-4; page 183

[10] Nielsen, Lykke M. Thesaurus construction: key issue and selected reading. // The thesaurus: review, renaissance, and revision / Roe ,Sandra K. ; and Thomas ,Alan R., (ed.). Binghamton: The Haworth Information Press, 2004., page 58

[11] Ibis, page 15-16.

[12] Bates, M.J. Task force recommendation 2.3 research and design review: improving user access to library catalog and portal information: final report (version 3). 2003. http://www.loc.gov/catdir/bibcontrol/2.3BatesReport6-03.doc.pdf (2009-09-18)

by the accuracy of data retrieved. It is the recall and the precision which attempt to measure the effectiveness.[13]

The information retrieval has changed dramatically in recent years, with the immense increase in availability of searchable full text and the increasing availability of powerful engines for searching the text. Today it is beginning to seem as if all information is available in full text.[14] However, there are many problems (ambiguity, synonyms, etc.) that indicate otherwise. Full text searching will always be valuable for browsing in any size of file but in large files, controlled language access searching will always support efficient retrieval.[15] Therefore, we can conclude that thesauri and indexing are required in facilitating information retrieval. Electronic thesaurus versions have strengthened its role as a search aid. Many operational systems accessible via the internet have incorporated thesauri in their interface as a part of their browsing and searching facilities.[16] Any of these types of system could produce better results by taking advantage of the presence of thesaurus. Most information professionals also point to the value-add of thesauri to justify the cost of traditional databases.[17]

Currently, most large-scale IR systems in general use consist of an indexed document database and a static thesaurus of terms and simple relationships. There are many such thesauri already in existence, designed in the first instance as printed documents to be consulted by human searchers, and there is an international standard setting out detailed rules for their compilation.[18]

An information retrieval thesaurus should, ideally, serve many purposes in information origination, storage and retrieval. Some of the more important applications of the thesaurus in such an environment are listed by Eugene Wall: 1) To serve as a term authority for indexers, so that only "acceptable" terms are employed by indexers. 2) To enable indexers quickly to find the "right" term to signify a concept in mind-"right" in the sense that the term must not only con-

[13] Muddamalle, Manikya Rao. Natural Language versus Controlled Vocabulary in Information Retrieval: A Case Study in Soil Mechanics. 1998. http://nlp.korea.ac.kr/new/seminar/2001spring/research/%5BMuddamalle98%5DNaturalLanguageVSControlledVocInIR.pdf (2009-07-22)

[14] Milstead, Jessica L. Use of Thesauri in the Full-Text Environment. 1998. http://www.bayside-indexing.com/Milstead/useof.htm (2009-07-22)

[15] Batty, David. WWW - Wealth, Weariness or Waste: Controlled Vocabulary and Thesauri in Support of Online Information Access. // D-Lib Magazine. 4 (1998), 10; page 2 http://www.dlib.org/dlib/november98/11batty.html (2009-07-22)

[16] Sihvonen, Anne; Vakkari, Pertti. Subject knowledge improves interactive query expansion assisted by a thesaurus. // Journal of Documentation. 60 (2004), 6; page 674

[17] Ojala, M. Finding and using the magic words: Keywords, thesauri and free text search. // Online. 31 (2007), 4; page 42

[18] Jones, Susan...[et al.]. Interactive thesaurus navigation: Intelligence rules OK? // Journal of the American Society for Information Science. 46 (1995), 1; page 53

note the proper concept but also must be appropriately specific (or general) with respect to the information being indexed. 3) To serve as a means of validating the results of the indexing effort, from the viewpoint of correctness of spelling, to insure that non-preferred synonyms are not employed by indexers, and to "flag" any terms newly required (in the indexer's judgment) by the system. 4) To enable the addition of cross-references between terms in any publication issue-periodic or cumulative-and to validate such cross-references to guarantee against circularity and ¨blindness¨ 5) To enable appropriate formulation of queries put to either printed or computerized indexes. 6) To provide a starting point for other systems which require a vocabulary significantly similar to the one encompassed by the thesaurus at hand. 7) To encourage consistent use of terminology by authors, abstractors, and other originators of information.[19]

ERIC database thesaurus

The Education Resources Information Center database is sponsored by the U.S. Department of Education to provide extensive access to educational-related literature. ERIC provides coverage of journal articles, conferences, meetings, government documents, theses, dissertations, reports, audiovisual media, bibliographies, directories, books and monographs. We can search ERIC using keywords or using descriptors from Thesaurus. Searching by keywords requires matching the exact words found in a record, while searching by descriptors allows location of the records indexed by subject, regardless of the terminology the author may have used.[20] The Thesaurus of ERIC Descriptors contains an alphabetical listing of terms used for indexing and searching in the ERIC database. This word-by-word alphabetical display provides a variety of information for each descriptor.[21] Except alphabetic search we can also use browsing the Thesaurus by 41 categories. Of course, there is a possibility of combining the selected descriptors with Boolean operators (basic and advanced search) to refine retrieval.[22]

Cross-references and relations between descriptors

Seven types of cross- references are used: Scope Note (SN), Use For (UF) and Use (USE) references, Narrower Terms (NT), Broader Terms (BT), Related Terms (RT) and Parenthetical Qualifiers.[23]

[19] Wall, Eugene. Symbiotic development of thesauri and information systems: A case history // Journal of the American Society for Information Science. 26 (1975), 2; pages 71-72

[20] ProQuest: ERIC. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

[21] ProQuest: ERIC Thesaurus. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

[22] ERIC: Education Resources Information Center: Search the Thesaurus. http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Thesaurus&_nfls=false (2009-08-01)

[23] Ibid.

Scope Note (SN)= brief statement of the intended usage of a descriptor. It may be used to clarify an ambiguous term or to restrict the usage of a term.[24]

Example:

INFORMATION RETRIEVAL

SN Techniques used to recover specific information from large quantities of stored data.[25]

Use For (UF) and *USE (USE)*= terms we consider to be equivalent (equal or almost equal by the meaning) we can combine to the category of equivalence so that equivalent expressions match only one term. Equivalence relations direct synonyms and pseudosynonyms of specific term to appropriate descriptor. For these relations we use UF and USE references.[26]
The UF reference is employed generally to solve problems of synonymy occurring in natural language. Terms following the UF notation are not used in indexing. They most often represent either (1) synonymous or variant forms of the main term, or (2) specific terms that, for purposes of storage and retrieval, are indexed under a more general term. Years listed in parentheses indicate the time period during which the term was used in indexing. It provides useful information for searching older printed indexes, or computer files that have not been updated.[27]

Example:

BIBLIOGRAPHIC DATABASES

UF Bibliographic Records (2004); Bibliographic Utilities (2004)[28]

The USE reference, the mandatory reciprocal of the UF, refers an indexer or searcher from a no usable (non indexable) term to the preferred indexable term or terms.[29]

[24] Craven, Tim. Thesaurus construction. 2008. http://publish.uwo.ca/~craven/677/thesaur/main00.htm. (2009-07-22)

[25] ERIC: Education Resources Information Center: Search the Thesaurus. http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Thesaurus&_nfls=false (2009-08-01)

[26] Urbanija, Jože. Ibid. Page 27

[27] ProQuest: ERIC Thesaurus. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

[28] ERIC: Education Resources Information Center: Search the Thesaurus. http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Thesaurus&_nfls=false (2009-08-01)

[29] ProQuest: ERIC Thesaurus. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

Example:

KINESCOPES
USE Films

Narrower Terms (NT) and *Broader Terms (BT)*= These indicate the existence of a hierarchical relationship between a class and its subclasses. In a hierarchical relation, one term is viewed as being "above" another term because it is broader in scope. Narrower terms are included in the broader class represented by the main entry. The Broader Term (BT) is the mandatory reciprocal of the NT. Broader Terms include as a subclass the concept represented by the main (narrower) term.[30]

Example:

SCHOOL CULTURE
BT Culture; Organizational Culture

Example:

RECREATIONAL ACTIVITIES
NT Playground Activities; Recreational Reading[31]

Related Terms (RT)= Associative relations express the analogy (not equivalence) between concepts. These kinds of relations are used for not hierarchical semantic relations in the thesaurus.[32]

Example:

ALCOHOLISM
RT Addictive Behaviour; Alcohol Education; Antisocial Behaviour; Behaviour Disorders; Drug Addiction; Fetal Alcohol Syndrome; Physical Health; Special Health Problems[33]

Parenthetical Qualifiers= A Parenthetical Qualifier is used to identify a particular indexable meaning of a homograph. In other words, it discriminates between terms (either Descriptors or USE references) that might otherwise be confused with each other. Examples include LETTERS (ALPHABET) and LETTERS (CORRESPONDENCE). The Qualifier is considered an integral part of

[30] Urbanija, Jože. Ibid. Page 28

[31] ERIC: Education Resources Information Center: Search the Thesaurus. http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Thesaurus&_nfls=false (2009-08-01)

[32] Urbanija, Jože. Ibid. Page 31

[33] ERIC: Education Resources Information Center: Search the Thesaurus. http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Thesaurus&_nfls=false (2009-08-01)

the Descriptor and must be used with the Descriptor in indexing and searching.[34]

LISTA database thesaurus

LISTA (Library, Information Science & Technology Abstracts) is a bibliographic database made available by EBSCO. The database offers searchable cited references, alerts functionality, author profiles and online tutorials. It provides coverage on subjects such as librarianship, classification, cataloguing, online information retrieval and information management. The thesauri in both the LISTA and LISTA with Full Text databases include 6,800 terms, 2,700 of which are preferred terms.[35] We can browse LISTA thesaurus by choosing tree type of displays: *Term begins with* displays a browsable alphabetical list, *Term contains* displays all the subject descriptors that contain requested term, whether it's the first word or not, and other terms to which requested term is related and *Relevancy ranked* displays the exact match to requested term first, if one exists, followed by subject terms "in order of relevance." As well as it is in the Thesaurus of ERIC Descriptors, here we can also combine two or more descriptors using Boolean operators to refine retrieval. Unlike the the Thesaurus of ERIC Descriptors in LISTA thesaurus there is no browsing by category list, just alphabetic list.. Another obvious difference from the Thesaurus of ERIC Descriptors is that cross-reference USE (USE) appears in all displays and there is possibility of ¨exploding¨ the term.[36]

Cross-references and relations between descriptors

Since we have explained types of relations between descriptors on the example of the Thesaurus of ERIC Descriptors, in the rest of the paper will be given only examples of that relations in LISTA thesaurus.

Six types of cross- references are used: Scope Note (SN), Use For (UF) and Use (USE) references, Narrower Terms (NT), Broader Terms (BT), and Related Terms (RT).

Scope Note (SN) example:

ACADEMIC librarians

SN Here are entered works on librarians who manage and maintain college and university libraries.

[34] ProQuest: ERIC Thesaurus. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

[35] EBSCO publishing: Customer Success Center: LISTA. 2009. http://www.ebscohost.com/customerSuccess/default.php?id=7 (2009-03-08)

[36] Library, Information Science & Technology Thesaurus. 2009. http://web.ebscohost.com/ehost/thesaurus?vid=2&hid=8&sid=bb81003c-cc48-4dfd-8a97-593c9d9ec7a8%40sessionmgr10 (2009-03-08)

Use (USE) example:
INFORMATION centres
USE INFORMATION services
Use For (UF) example:
PUBLIC domain (Copyright law)
UF COPYRIGHT-- Public domain
Narrower Term (NT) example:
COMPUTER FILES
NT COMPUTER programs; DATABASES; IMAGE files; TEXT files
Broader Terms (BT) example:
TELEGRAPH
BT TELECOMMUNICATION
Related Terms (RT) example:
SCHOLARLY publishing
RT ACADEMIC writing; CONFERENCE proceedings; MONOGRAPHIC series; SCHOLARLY periodicals; UNIVERSITY presses[37]

Conclusion

Today, when information sources are growing enormously, there is a need for more effective information retrieval. Although in each database we have possibility to use Boolean operators to refine our retrieval it seems that is not enough. This is because of linguistic problems that can occur. Thesaurus copes with these problems very well so we can conclude that this tool is vital retrieval tool in databases. The main problem that remains is limited users' thesauri comprehension. If we could correct and overhaul these problems thesauri would probably be more useful for users during IR processes.

References

ANSI/NISO Z39.19-2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. 2005. http://eric.de.gov/ERICWebPortal/resources/html/help/Z39-19-2005.pdf (2009-07-22)

Bates, M.J. Task force recommendation 2.3 research and design rewiew: improving user acess to library catalog and portal information: final report (version 3). 2003. http://www.loc.gov/catdir/bibcontrol/2.3BatesReport6-03.doc.pdf (2009-09-18)

Batty, David. WWW - Wealth, Weariness or Waste: Controlled Vocabulary and Thesauri in Support of Online Information Access. // *D-Lib Magazine*. 4 (1998), 10; pages 1-6 www.dlib.org/dlib/november98/11batty.html (2009-07-22)

Bawden, David. Tezaurusi: nova postignuća. // *Vjesnik bibliotekara Hrvatske*. 44 (2001), 1-4; pages 182-187

Bibliotekarski leksikon. Beograd: Nolit, 1984.

Craven, Tim. Thesaurus construction. 2008. http://publish.uwo.ca/~craven/677/thesaur/main00.htm. (2009-07-22)

[37] Ibid.

EBSCO publishing: Customer Success Center: LISTA. 2009. http://www.ebscohost.com/customerSuccess/default.php?id=7 (2009-03-08)

ERIC: Education Resources Information Center: Search the Thesaurus. http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Thesaurus&_nfls=false (2009-08-01)

Greenberg, J. User comprehension and searching with information retrieval thesauri // The thesaurus: review, renaissance, and revision / Roe ,Sandra K. ; and Thomas ,Alan R., (ed.). Binghamton: The Haworth Information Press, 2004., page 103-120.

Jones, Susan.[et al.]. Interactive thesaurus navigation: Intelligence rules OK? // Journal of the American Society for Information Science. 46 (1995), 1; pages 52-59

Lancaster, F. W. Do indexing and abstracting have a future? // Anales de documentation. 2003, 6; pages 137-144

Lancaster, F.W. Indexing and abstracting in theory and practice. Compaing Illions: University of Illinois, 1998.

Leščić, Jelica. O tezaurusu načela, izradba, struktura: pregled. // Vjesnik bibliotekara Hrvatske 44 (2001), 1-4; pages 172-181

Library, Information Science & Technology Thesaurus. 2009. http://web.ebscohost.com/ehost/thesaurus?vid=2&hid=8&sid=bb81003c-cc48-4dfd-8a97-593c9d9ec7a8%40sessionmgr10 (2009-03-08)

Milstead, Jessica L. Use of Thesauri in the Full-Text Environment. 1998. http://www.bayside-indexing.com/Milstead/useof.htm (2009-07-22)

Muddamalle, Manikya Rao. Natural Language versus Controlled Vocabulary in Information Retrieval: A Case Study in Soil Mechanics. 1998. http://nlp.korea.ac.kr/new/seminar/2001spring/research/%5BMuddamalle98%5DNaturalLanguageVSControlledVocInIR.pdf (2009-07-22)

Nielsen, Lykke M. Thesaurus construction: key issue and selected reading. // The thesaurus: review, renaissance, and revision / Roe ,Sandra K. ; and Thomas ,Alan R., (ed.). Binghamton: The Haworth Information Press, 2004., pages 57-74

Ojala, M. Finding and using the magic words: Keywords, thesauri and free text search. // *Online*. 31 (2007), 4; pages 40-42

ProQuest: ERIC. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

ProQuest: ERIC Thesaurus. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

Sihvonen, Anne; Vakkari, Pertti. Subject knowledge improves interactive query expansion assisted by a thesaurus. // *Journal of Documentation*. 60 (2004), 6; pages 673-690

Svenonius, Elaine. Intelektualne osnove organizacije informacija. Lokve: Benja, 2005.

Urbanija, Jože. Metodologija izrade tezaurusa. Zagreb: Naklada Nediljko Dominović, 2005.

Wall, Eugene. Symbiotic development of thesauri and information systems: A case history // *Journal of the American Society for Information Science*. 26 (1975), 2; pages 71-79